Comparison of machine learning models for water quality prediction: a case study of Lake Daecheong, Korea

H. Sim[1,2], J. Yoon[1,2,3*] & C. E. Han[2,4,*]

[1]*Program in Environmental Technology and Policy, Korea University, Sejong 30019, Korea*

[2]*Interdisciplinary Graduate Program in Artificial Intelligence Technology for Smart Convergence, Korea University, Sejong 30019, Korea*

[3]*Department of Environmental Engineering, Korea University, Sejong 30019, Korea*

[4]*Department of Electronics and Information Engineering, Korea University, Sejong 30019, Korea*

**Corresponding authors' email: jyyoon@korea.ac.kr; cheolhan@korea.ac.kr*

Highlights

- Four machine learning techniques were compared for their water quality prediction capability.
- DO and BOD were the parameters that were predicted best, and SS was modelled worst.
- MLR worked best for DO, BOD and SS; KNN for COD and TN; and RF for TP.

Introduction

Due to population growth and industrial development, water use and associated water pollution have increased. Accordingly, an active response is being made to mitigate the pollution of receiving waters such as lakes. In general, a water quality prediction method used for management comes in the form of a numerical model based on the mass-balance equation. However, data necessary for the estimation of relevant parameters are lacking. Therefore, many studies on water quality prediction have turned to machine learning techniques, which have more flexible structure and fewer parameters to be estimated (Ahmed et al., 2019; Chou et al., 2018). Though successful, they either did not cover a popular set of water quality parameters such as BOD and COD or did not test commonly available machine learning methods such as multiple linear regression and support vector machine regression. In this study, four commonly available machine learning techniques were used to predict the popular water quality parameters of a lake. The feasibility of machine learning based water quality prediction was then investigated by identifying models with good efficiency.

Methodology

Study site

In this study, Lake Daecheong (Figure 1), a main water supply source for central western region of South Korea, was selected as the study site, located 150 km from the mouth of the Geum River. The Geum River is 394.79 km long and has 9,912.15 km^2 of watershed area, which is the third largest in South Korea.

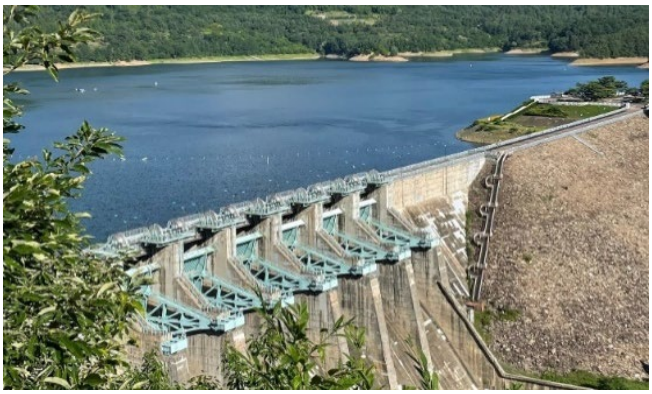

Figure 1. Lake Daecheong

Data

The water environment data of lakes publicly provided by the Water Environment Information System was used, and the data collection period is from January 2007 to October 2020 (The Water Environment Information System website, 2021). Selected data are monthly values of water quality parameters such as dissolved oxygen (DO) [mg/L], biochemical oxygen demand (BOD) [mg/L], chemical oxygen demand (COD) [mg/L], suspended solids (SS) [mg/L], total nitrogen (TN) [mg/L], total phosphorus (TP) [mg/L]; environmental parameters such as pH [-], water temperature [°C] and electrical conductivity [µs/cm]; and hydrological parameters such as water storage [m^3] and rainfall [mm].

Machine Learning methods

Machine learning is one of the research fields of artificial intelligence, and it is a method that aims to realize functions such as human-style learning in computers. In this study, four popular machine learning techniques were used, namely, multiple linear regression (MLR), support vector machine regression (SVMR) (Drucker et al., 1997; Kavitha et al., 2016), K-nearest neighbor (KNN) (Cover, 1968), and random forest (RF) (Ho, 1995; Rodriguez-Galiano et al., 2015); and we compared their applicability for water quality prediction.

We constructed models to predict each of water quality parameters as dependent variables using environmental and hydrological parameters as independent variables. A water quality parameter for a particular month was predicted based on the environmental parameters of the same month, and hydrological parameters of the same month and one and two months ago to account for their influences. We standardized the independent variables for SVMR, KNN, and RF to put them on the same scale. The data was split into training and test sets with 8:2 ratio. We found optimal hyper-parameters of each model through 5-fold cross-validation over the training set. To implement the models, we used Python and scikit-learn (Pedregosa et al., 2011). We used in-house codes for MLR, and we adapted the codes developed by Hong (2021) to suit our needs for SVMR, KNN and RF.

Model performance evaluation

We utilized the two model performance statistics to identify the best machine learning models for water quality prediction: coefficient of determination (R^2) and root mean square error (RMSE). R^2 was primarily used to discern the relative performance across the different water quality parameters while RMSE was referred to as an auxiliary measure of accuracy. Since different water quality parameters have different range of magnitudes, RMSE, which carries the unit of a water quality parameter, is difficult to be used as a comparative measure. On the other hand, R^2 is a normalized metric, and thus there is no such problems, where the closer value to 1 represents the better applicability.

Results and discussion

Water quality simulation results for each machine learning method are presented in Table 1.

Table 1. Performance summary of each machine learning model for water quality parameters

Water quality parameter (mg/L)	Evaluation index	MLR	SVMR	KNN	RF
Dissolved Oxygen (DO)	R^2	**0.88**	0.66	0.8	0.87
	RMSE	**0.9**	2.23	1.12	0.91
Biochemical Oxygen Demand (BOD)	R^2	**0.79**	0.7	0.65	0.7
	RMSE	0.11	0.02	0.02	**0.01**
Chemical Oxygen Demand (COD)	R^2	0.53	0.4	**0.61**	0.41
	RMSE	0.42	0.35	**0.15**	0.41
Suspended Solids (SS)	R^2	**0.21**	0.17	0.07	0.3
	RMSE	**1.2**	2.69	4.3	2.3
Total Nitrogen (TN)	R^2	0.2	0.07	**0.56**	0.35
	RMSE	0.28	0.07	**0**	0.05
Total Phosphorus (TP)	R^2	0.51	-0.98	0.23	**0.54**
	RMSE	0.01	0.0	0.05	**0**

Best values for each water quality parameter were marked in bold.

For a proper evaluation of model performance, we borrowed the criteria suggested by Moriasi (2015) for water quality parameters where in general R^2 greater than 0.6 was considered good, between 0.3 and 0.6 satisfactory and less than 0.3 unsatisfactory. According to these criteria, of the water quality parameters considered, DO, BOD and COD were identified to be the ones that are modelled good by at least one of the machine learning methods evaluated. Followed by these three parameters, TN and TP were further identified to be the ones that can be modelled satisfactorily. Finally, SS was found to be the parameter that was modelled worst.

Of the machine learning methods evaluated, the best performer was not always the same for all water quality parameters. MLR gave the best performance for DO, BOD, and SS. KNN and RF gave the best performance for COD and TP, respectively. Lastly, SVMR was the least performing method that did not result in the best performance for any of the water quality parameters considered. We note that MLR is a rather statistical method which assumes linearity between independent and dependent variables. In that sense, high performances by MLR on DO and BOD imply their linear relationship to dependent variables considered. For the parameters where MLR was not the best performer, such as COD, TN and TP, we can expect the linear relationship was not as strong as DO and BOD, since some of the machine learning models utilizing nonlinearity of the data performed better.

Overall, it was found that some water quality parameters can be modelled satisfactorily by the machine learning methods. However, their applicability can be different for different water quality parameters.

Conclusions and future work

In this study, the feasibility of predicting water quality of Lake Daecheong was evaluated using four machine learning techniques. Overall satisfactory results were obtained for most water quality parameters except SS. This shows the promise of machine learning methods in predicting water quality parameters of a lake which can then be usefully utilized for subsequent water quality management. While viewed as encouraging results, it is judged that there is also a limitation to water quality prediction based only on water temperature, electrical conductivity, pH, rainfall, and water storage. And predictive application of this sort should have been benefited from a larger number of data if they were available. Therefore, investigation of different combinations of independent variables, including other factors than the ones considered in this study, for existing regression analysis methods and possibly the utilization of new methods such as deep learning techniques along with growing number of datasets, are recommended as future works that can contribute to the advancement of water quality prediction technology for lakes.

References

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M. and Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084.

Chou, J. S., Ho, C. C., and Hoang, H. S. (2018). Determining quality of water in reservoir using machine learning. Ecological informatics, 44, 57-75.

Cover, T. (1968). Estimation by the nearest neighbor rule. IEEE Transactions on Information Theory, 14(1), 50-55.

Drucker, H., Burges, C. C., Kaufman, L., Smola, A. J., and Vapnik, V. N. (1997). Support Vector Regression Machines. in Advances in Neural Information Processing Systems 9, NIPS 1996, 155–161.

Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282.

Hong, S. B. K-nearest neighbor & support vector machine regression & random forest. https://github.com/baek2sm/MLCook/tree/master/sklearn-cook/regression (Accessed in June 2021).

Kavitha, S., Varuna, S., and Ramya, R. (2016). A comparative analysis on linear regression and support vector regression. In 2016 online international conference on green engineering and technologies (IC-GET). IEEE, 1-5.

Moriasi, D. N., Gitau, M. W., Pai, N. and Daggupati, P. (2015). Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. American Society of Agricultural and Biological Engineers, 58(6), 1763-1785.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.

The Water Environment Information System website. http://water.nier.go.kr/main/mainContent.do (Accessed in June 2021).